Proceedings of Machine Learning Research vol 98:1–23, 2019 30th International Conference on Algorithmic Learning Theory

Sample Compression for Real-Valued Learners

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Toyota Technological Institute at Chicago

Aryeh Kontorovich KARYEH@CS.BGU.AC.IL
Ben-Gurion University

Menachem Sadigurschi SADIGURS@POST.BGU.AC.IL
Ben-Gurion University

Editors: Aurélien Garivier and Satyen Kale

Abstract

We give an algorithmically efficient version of the learner-to-compression scheme conversion in Moran and Yehudayoff (2016). We further extend this technique to real-valued hypotheses, to obtain a bounded-size sample compression scheme via an efficient reduction to a certain generic real-valued learning strategy. To our knowledge, this is the first general compressed regression result (regardless of efficiency or boundedness) guaranteeing uniform approximate reconstruction. Along the way, we develop a generic procedure for constructing weak real-valued learners out of abstract regressors; this result is also of independent interest. In particular, this result sheds new light on an open question of H. Simon (1997). We show applications to two regression problems: learning Lipschitz and bounded-variation functions.

Keywords: Compression Scheme, Boosting, Regression, Empirical Risk Minimization

1. Introduction

Sample compression is a natural learning strategy, whereby the learner seeks to retain a small subset of the training examples, which (if successful) may then be decoded as a hypothesis with low empirical error. Overfitting is controlled by the size of this learner-selected "compression set". Part of a more general *Occam learning* paradigm, such results are commonly summarized by "compression implies learning". A fundamental question, posed by Littlestone and Warmuth (1986), concerns the reverse implication: Can every learner be converted into a sample compression scheme? Or, in a more quantitative formulation: Does every VC class admit a constant-size sample compression scheme? A series of partial results (Floyd, 1989; Helmbold et al., 1992; Floyd and Warmuth, 1995; Ben-David and Litman, 1998; Kuzmin and Warmuth, 2007; Rubinstein et al., 2009; Rubinstein and Rubinstein, 2012; Chernikov and Simon, 2013; Livni and Simon, 2013; Moran et al., 2017) culminated in Moran and Yehudayoff (2016) which resolved the latter question[1].

1. The refined conjecture of Littlestone and Warmuth (1986), that any concept class with VC-dimension d admits a compression scheme of size $O(d)$, remains open.

Moran and Yehudayoff's solution involved a clever use of von Neumann's minimax theorem, which allows one to make the leap from the existence of a weak learner uniformly over all *distributions on examples* to the existence of a *distribution on weak hypotheses* under which they achieve a certain performance simultaneously over all of the examples. Although their paper can be understood without any knowledge of boosting, Moran and Yehudayoff note the well-known connection between boosting and compression. Indeed, boosting may be used to obtain a constructive proof of the minimax theorem (Freund and Schapire, 1996, 1999) — and this connection was what motivated us to seek an efficient algorithm implementing Moran and Yehudayoff's existence proof. Having obtained an efficient conversion procedure from consistent PAC learners to bounded-size sample compression schemes, we turned our attention to the case of real-valued hypotheses, seeking to apply this same approach. In this case, it turned out that getting this approach to yield bounded compression schemes required significant innovation in the technical details of the proof. In particular, for the boosting approach, there are several different notions of "weak learner" that could be considered. It turns out one such definition is appropriate for the purpose of compression, while the others are not. This then leads to additional questions, as unlike the binary case, there was not already in the literature an understanding of the sample complexity of this notion of weak learning, which is an important part of the analysis of the size of the compression scheme. We therefore needed to supply such an analysis. Finally, a critical component in the compression approach for classification is a *sparsification* step, which is the main innovation that enabled Moran and Yehudayoff to remove the dependence on the data set size from the size of the compression set. This step also required a significantly different technique in the proof to arrive at such a sparse subset in the case of real-valued functions. Nevertheless, with all of these components established, we were indeed able to construct a bounded-size compression scheme for classes of real-valued functions, following the same high-level strategy from the binary-valued case.

Our contribution. More formally, in the classification setting, our technique combines the innovations of Moran and Yehudayoff (2016), with the simple but powerful observation (Schapire and Freund, 2012) that many boosting algorithms (e.g., AdaBoost, α-Boost) are capable of outputting a family of $O(\log(m)/\gamma^2)$ hypotheses such that not only does their (weighted) majority vote yield a sample-consistent classifier, but in fact a $\approx (\frac{1}{2}+\gamma)$ super-majority does as well. This fact implies that after boosting, we can sub-sample a constant (i.e., independent of sample size m) number of classifiers and thereby efficiently recover the sample compression bounds of Moran and Yehudayoff (2016).

But our chief technical contribution is in the real-valued case. As we discuss below, extending the boosting framework from classification to regression presents a host of technical challenges. One of our insights is to impose distinct error metrics on the weak and strong learners: a "stronger" one on the latter and a "weaker" one on the former. This allows us to achieve two goals simultaneously:

(a) We give apparently the first generic analysis of the sample complexity of weak (and strong) learning (in the sense defined below) for real-valued functions, via simple sample-consistent learning rules. This is in contrast with many previous proposed weak regressors, whose stringent or exotic definitions made them unwieldy to construct or verify as such, and most of which are not compatible with generic weak-to-strong boosting strategies. This result is novel and is also of independent interest.

(b) We show that the output of a certain real-valued boosting algorithm may be sparsified so as to yield a constant size sample compression scheme: a real-valued analogue of the Moran and Yehudayoff result for classification. This gives the first general constant-size sample compression scheme having uniform approximation guarantees on the data.

2. Definitions and notation

We will write $[k] := \{1, \ldots, k\}$. An *instance space* is an abstract set $\mathcal{X}$. For a concept class $\mathcal{C} \subset \{0,1\}^{\mathcal{X}}$, if say that $\mathcal{C}$ *shatters* a set $\{x_1, \ldots, x_k\} \subset \mathcal{X}$ if

$$\mathcal{C}(S) = \{(f(x_1), f(x_2), \ldots, f(x_k)) : f \in \mathcal{C}\} = \{0,1\}^k .$$

The VC-dimension $d = d_{\mathcal{C}}$ of $\mathcal{C}$ is the size of the largest shattered set (or ∞ if $\mathcal{C}$ shatters sets of arbitrary size) (Vapnik and Červonenkis, 1971). When the roles of $\mathcal{X}$ and $\mathcal{C}$ are exchanged — that is, an $x \in \mathcal{X}$ acts on $f \in \mathcal{C}$ via $x(f) = f(x)$, — we refer to $\mathcal{X} = \mathcal{C}^*$ as the *dual* class of $\mathcal{C}$. Its VC-dimension is then $d^* = d^*_{\mathcal{C}} := d_{\mathcal{C}^*}$, and referred to as the *dual VC dimension*. Assouad (1983) showed that $d^* \le 2^{d+1}$.

For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ and $t > 0$, we say that $\mathcal{F}$ t-shatters a set $\{x_1, \ldots, x_k\} \subset \mathcal{X}$ if there is an $r \in \mathbb{R}^m$ such that for all $y \in \{-1,1\}^m$ there is an $f \in \mathcal{F}$ such that $\min_{i \in [k]} y_i(f(x_i) - r_i) \ge t$. The t-fat-shattering dimension $d(t) = d_{\mathcal{F}}(t)$ is the size of the largest t-shattered set (possibly ∞) (Alon et al., 1997). Again, the roles of $\mathcal{X}$ and $\mathcal{F}$ may be switched, in which case $\mathcal{X} = \mathcal{F}^*$ becomes the dual class of $\mathcal{F}$. Its t-fat-shattering dimension is then $d^*(t)$, and Assouad's argument shows that $d^*(t) \le 2^{d(t)+1}$.

A *sample compression scheme* (κ, ρ) for a hypothesis class $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ is defined as follows. A k-*compression* function κ maps sequences $((x_1, y_1), \ldots, (x_m, y_m)) \in \bigcup_{\ell \ge 1} (\mathcal{X} \times \mathcal{Y})^\ell$ to elements in $\mathcal{K} = \bigcup_{\ell \le k'} (\mathcal{X} \times \mathcal{Y})^\ell \times \bigcup_{\ell \le k''} \{0,1\}^\ell$, where $k' + k'' \le k$. A *reconstruction* is a function $\rho : \mathcal{K} \to \mathcal{Y}^{\mathcal{X}}$. We say that (κ, ρ) is a k-size sample compression scheme for $\mathcal{F}$ if κ is a k-compression and for all $h^* \in \mathcal{F}$ and all $S = ((x_1, h^*(x_1)), \ldots, (x_m, h^*(y_m)))$, we have $\hat{h} := \rho(\kappa(S))$ satisfies $\hat{h}(x_i) = h^*(x_i)$ for all $i \in [m]$.

For real-valued functions, we say it is a *uniformly* ε-*approximate* compression scheme if

$$\max_{1 \le i \le m} |\hat{h}(x_i) - h^*(x_i)| \le \varepsilon.$$

3. Main results

Throughout the paper, we implicitly assume that all hypothesis classes are *admissible* in the sense of satisfying mild measure-theoretic conditions, such as those specified in Dudley (1984, Section 10.3.1) or Pollard (1984, Appendix C). This section states our main results. The remainder of the paper is dedicated to presenting their proofs. We begin with an algorithmically efficient version of the learner-to-compression scheme conversion in Moran and Yehudayoff (2016):

Theorem 1 (Efficient compression for classification) *Let $\mathcal{C}$ be a concept class over some instance space $\mathcal{X}$ with VC-dimension d, dual VC-dimension d^*, and suppose that $\mathcal{A}$ is a (proper, consistent) PAC-learner for $\mathcal{C}$: For all $0 < \varepsilon, \delta < 1/2$, all $f^* \in \mathcal{C}$, and all distributions D over $\mathcal{X}$, if $\mathcal{A}$ receives $m \geq m_{\mathcal{C}}(\varepsilon, \delta)$ points $S = \{x_i\}$ drawn iid from D and labeled with $y_i = f^*(x_i)$, then $\mathcal{A}$ outputs an $\hat{f} \in \mathcal{C}$ such that*

$$\mathbb{P}_{S\sim D^m}\left(\mathbb{P}_{X\sim D}\left(\hat{f}(X) \neq f^*(X) \,|\, S\right) > \varepsilon\right) < \delta.$$

For every such $\mathcal{A}$, there is a randomized sample compression scheme for $\mathcal{C}$ of size $O(k \log k)$, where $k = O(dd^*)$. Furthermore, on a sample of any size m, the compression set may be computed in expected time

$$O\left((m + T_{\mathcal{A}}(cd)) \log m + mT_{\mathcal{E}}(cd)(d^* + \log m)\right),$$

where $T_{\mathcal{A}}(\ell)$ is the runtime of $\mathcal{A}$ to compute $\hat{f}$ on a sample of size ℓ, $T_{\mathcal{E}}(\ell)$ is the runtime required to evaluate $\hat{f}$ on a single $x \in \mathcal{X}$, and c is a universal constant.

Although for our purposes the existence of a distribution-free sample complexity $m_{\mathcal{C}}$ is more important than its concrete form, we may take $m_{\mathcal{C}}(\varepsilon, \delta) = O(\frac{d}{\varepsilon} \log \frac{1}{\varepsilon} + \frac{1}{\varepsilon} \log \frac{1}{\delta})$ (Vapnik and Chervonenkis, 1974; Blumer et al., 1989), known to bound the sample complexity of empirical risk minimization; indeed, this loses no generality, as there is a well-known efficient reduction from empirical risk minimization to any proper learner having a polynomial sample complexity (Pitt and Valiant, 1988; Haussler et al., 1991). We allow the evaluation time of $\hat{f}$ to depend on the size of the training sample in order to account for non-parametric learners, such as nearest-neighbor classifiers. A naive implementation of the Moran and Yehudayoff (2016) existence proof yields a runtime of order $m^{cd} T_{\mathcal{A}}(c'd) + m^{cd^*}$ (for some universal constants c, c'), which can be doubly exponential when $d^* = 2^d$; this is without taking into account the cost of computing the minimax distribution on the $m^{cd} \times m$ game matrix.

Next, we extend the result in Theorem 1 from classification to regression:

Theorem 2 (Efficient compression for regression) *Let $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ be a function class with t-fat-shattering dimension $d(t)$, dual t-fat-shattering dimension $d^*(t)$, and suppose that $\mathcal{A}$ is an ERM (i.e., proper, consistent) learner for $\mathcal{F}$: For all $f^* \in \mathcal{C}$, and all distributions D over $\mathcal{X}$, if $\mathcal{A}$ receives m points $S = \{x_i\}$ drawn iid from D and labeled with $y_i = f^*(x_i)$, then $\mathcal{A}$ outputs an $\hat{f} \in \mathcal{F}$ such that $\sum_{i\in[m]} |\hat{f}(x_i) - f^*(x_i)| = 0$. For every such $\mathcal{A}$, there is a randomized uniformly ε-approximate sample compression scheme for $\mathcal{F}$ of size $O(k\tilde{m} \log(k\tilde{m}))$, where $\tilde{m} = O(d(c\varepsilon) \log(1/\varepsilon))$ and $k = O(d^*(c\varepsilon) \log(d^*(c\varepsilon)/\varepsilon))$. Furthermore, on a sample of any size m, the compression set may be computed in expected time*

$$O(mT_{\mathcal{E}}(\tilde{m})(k + \log m) + T_{\mathcal{A}}(\tilde{m}) \log(m)),$$

where $T_{\mathcal{A}}(\ell)$ is the runtime of $\mathcal{A}$ to compute $\hat{f}$ on a sample of size ℓ, $T_{\mathcal{E}}(\ell)$ is the runtime required to evaluate $\hat{f}$ on a single $x \in \mathcal{X}$, and c is a universal constant.

A key component in the above result is our construction of a generic (η, γ)-weak learner.

Definition 3 *For $\eta \in [0,1]$ and $\gamma \in [0,1/2]$, we say that $f : \mathcal{X} \to \mathbb{R}$ is an an (η,γ)-*weak hypothesis *(with respect to distribution D and target $f^* \in \mathcal{F}$) if*

$$\mathbb{P}_{X\sim D}(|f(X) - f^*(X)| > \eta) \leq \frac{1}{2} - \gamma.$$

Theorem 4 (Generic weak learner) *Let $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ be a function class with t-fat-shattering dimension $d(t)$. For some universal numerical constants $c_1, c_2, c_3 \in (0,\infty)$, for any $\eta, \delta \in (0,1)$ and $\gamma \in (0,1/4]$, any $f^* \in \mathcal{F}$, and any distribution D, letting $X_1, \dots, X_m$ be drawn iid from D, where*

$$m = \left\lceil c_1 \left(d(c_2\eta) \ln\left(\frac{c_3}{\eta}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right\rceil,$$

with probability at least $1-\delta$, every $f \in \mathcal{F}$ with $\sum_{i\in[m]} |f(X_i) - f^*(X_i)| = 0$ is an (η,γ)-weak hypothesis with respect to D and f^*.

Remark: In fact, our results would also allow us to use any hypothesis $f \in \mathcal{F}$ with $\max_{i\in[m]} |f(X_i) - f^*(X_i)|$ merely smaller than η by a constant factor: for instance, bounded by $\eta/2$. This can then also be plugged into the construction of the compression scheme and this criterion can be used in place of consistency in Theorem 2. This also enables our compression scheme to be applied in settings that are not strictly realizable, but rather have achievable ℓ_∞ loss at most η/c for some $c > 1$.

In Sections B and A we give applications to sample compression for nearest-neighbor and bounded-variation regression.

4. Related work

It appears that generalization bounds based on sample compression were independently discovered by Littlestone and Warmuth (1986) and Devroye et al. (1996) and further elaborated upon by Graepel et al. (2005); see Floyd and Warmuth (1995) for background and discussion. A more general kind of Occam learning was discussed in Blumer et al. (1989). Computational lower bounds on sample compression were obtained in Gottlieb et al. (2014), and some communication-based lower bounds were given in Kane et al. (2017).

Beginning with Freund and Schapire (1997)'s `AdaBoost.R` algorithm, there have been numerous attempts to extend AdaBoost to the real-valued case (Bertoni et al., 1997; Drucker, 1997; Avnimelech and Intrator, 1999; Karakoulas and Shawe-Taylor, 2000; Duffy and Helmbold, 2002; Kégl, 2003; Nock and Nielsen, 2007) along with various theoretical and heuristic constructions of particular weak regressors (Mason et al., 1999; Friedman, 2001; Mannor and Meir, 2002); see also the survey Mendes-Moreira et al. (2012).

Duffy and Helmbold (2002, Remark 2.1) spell out a central technical challenge: no boosting algorithm can "always force the base regressor to output a useful function by simply modifying the distribution over the sample". This is because unlike a binary classifier, which localizes errors on specific examples, a real-valued hypothesis can spread its error evenly over the entire sample, and it will not be affected by reweighting. The (η,γ)-weak learner, which has appeared, among other works, in Anthony et al. (1996); Simon (1997); Avnimelech and Intrator (1999); Kégl (2003), gets around this difficulty — but provable general constructions of such learners have been lacking. Likewise, the heart of our sample compression engine, `MedBoost`, has been widely in use since Freund and Schapire (1997) in various guises. Our Theorem 4 supplies the remaining piece of the puzzle: *any* sample-consistent regressor applied to some random sample of bounded size yields an (η, γ)-weak hypothesis. The closest analogue we were able to find was Anthony et al. (1996, Theorem 3), which is non-trivial only for function classes with finite pseudo-dimension, and is inapplicable, e.g., to classes of 1-Lipschitz or bounded variation functions.

The literature on general sample compression schemes for real-valued functions is quite sparse. There are well-known narrowly tailored results on specifying functions or approximate versions of functions using a finite number of points, such as the classical fact that a polynomial of degree p can be perfectly recovered from $p+1$ points. To our knowledge, the only *general* results on sample compression for real-valued functions (applicable to *all* learnable function classes) is Theorem 4.3 of David, Moran, and Yehudayoff (2016). They propose a general technique to convert any learning algorithm achieving an arbitrary sample complexity $M(\varepsilon, \delta)$ into a compression scheme of size $O(M(\varepsilon, \delta) \log(M(\varepsilon, \delta)))$, where δ may approach 1. However, their notion of compression scheme is significantly weaker than ours: namely, they allow $\hat{h} = \rho(\kappa(S))$ to satisfy merely $\frac{1}{m}\sum_{i=1}^{m} |\hat{h}(x_i) - h^*(x_i)| \leq \varepsilon$, rather than our *uniform* ε-approximation requirement $\max_{1\leq i\leq m} |\hat{h}(x_i) - h^*(x_i)| \leq \varepsilon$. In particular, in the special case of $\mathcal{F}$ a family of *binary*-valued functions, their notion of sample compression does *not* recover the usual notion of sample compression schemes for classification, whereas our uniform ε-approximate compression notion *does* recover it as a special case. We therefore consider our notion to be a more fitting generalization of the definition of sample compression to the real-valued realizable (or nearly-realizable) case. On the other hand, in a sibling paper to the present work, we explore the subject of *agnostic-case* sample compression schemes for real-valued functions (Hanneke, Kontorovich, and Sadigurschi, 2018). In that work, we find that the definition of compression scheme studied by David, Moran, and Yehudayoff (2016) is most appropriate for the agnostic case, due to a strong connection to the generalization ability of the corresponding learning algorithm. Under that definition, that work constructs bounded-size sample compression schemes for agnostic learning of linear functions under ℓ_1 and ℓ_∞ losses, and further argues that these are the only ℓ_p losses for which such bounded-size compression schemes exist. It also poses a general question about the existence of bounded-size agnostic compression schemes for arbitrary classes of finite pseudo-dimension under ℓ_1 loss.

5. Boosting Real-Valued Functions

As mentioned above, the notion of a *weak learner* for learning real-valued functions must be formulated carefully. The naïve thought that we could take any learner guaranteeing, say, absolute loss at most $\frac{1}{2} - \gamma$ is known to not be strong enough to enable boosting to ε loss. However, if we make the requirement too strong, such as in Freund and Schapire (1997) for `AdaBoost.R`, then the sample complexity of weak learning will be so high that weak learners cannot be expected to exist for large classes of functions. However, our Definition 3, which has been proposed independently by Simon (1997) and Kégl (2003), appears to yield the appropriate notion of *weak learner* for boosting real-valued functions.

In the context of boosting for real-valued functions, the notion of an (η, γ)-weak hypothesis plays a role analogous to the usual notion of a weak hypothesis in boosting for classification. Specifically, the following boosting algorithm was proposed by Kégl (2003). As it will be convenient for our later results, we express its output as a sequence of functions and weights; the boosting guarantee from Kégl (2003) applies to the weighted quantiles (and in particular, the weighted median) of these function values.

Algorithm 1: $\texttt{MedBoost}(\{(x_i, y_i)\}_{i\in[m]}, T, \gamma, \eta)$

Define P_0 as the uniform distribution over $\{1, \ldots, m\}$
for $t = 0, \ldots, T$ **do**
  Call weak learner to get h_t and $(\eta/2, \gamma)$-weak hypothesis wrt $(x_i, y_i) : i \sim P_t$ (repeat until it succeeds)
  for $i = 1, \ldots, m$ **do**
    $\theta_i^{(t)} \leftarrow 1 - 2\mathbb{I}[|h_t(x_i) - y_i| > \eta/2]$
  end for
  $\alpha_t \leftarrow \frac{1}{2}\ln\left(\frac{(1-\gamma)\sum_{i=1}^m P_t(i)\mathbb{I}[\theta_i^{(t)}=1]}{(1+\gamma)\sum_{i=1}^m P_t(i)\mathbb{I}[\theta_i^{(t)}=-1]}\right)$
  if $\alpha_t = \infty$ **then**
    Return T copies of h_t, and $(1, \ldots, 1)$
  end if
  for $i = 1, \ldots, m$ **do**
    $P_{t+1}(i) \leftarrow P_t(i)\frac{\exp\{-\alpha_t\theta_i^{(t)}\}}{\sum_{j=1}^m P_t(j)\exp\{-\alpha_t\theta_j^{(t)}\}}$
  end for
end for
Return $(h_1, \ldots, h_T)$ and $(\alpha_1, \ldots, \alpha_T)$

Here we define the weighted median as

$$\mathrm{Median}(y_1, \ldots, y_T; \alpha_1, \ldots, \alpha_T) = \min\left\{y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j < y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2}\right\}.$$

Also define the weighted *quantiles*, for $\gamma \in [0, 1/2]$, as

$$Q_\gamma^+(y_1, \ldots, y_T; \alpha_1, \ldots, \alpha_T) = \min\left\{y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j < y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} - \gamma\right\}$$
$$Q_\gamma^-(y_1, \ldots, y_T; \alpha_1, \ldots, \alpha_T) = \max\left\{y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j > y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} - \gamma\right\},$$

and abbreviate $Q_\gamma^+(x) = Q_\gamma^+(h_1(x), \ldots, h_T(x); \alpha_1, \ldots, \alpha_T)$ and $Q_\gamma^-(x) = Q_\gamma^-(h_1(x), \ldots, h_T(x); \alpha_1, \ldots, \alpha_T)$ for $h_1, \ldots, h_T$ and $\alpha_1, \ldots, \alpha_T$ the values returned by `MedBoost`.

Then Kégl (2003) proves the following result.

Lemma 5 *(Kégl (2003)) For a training set $Z = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of size m, the return values of* `MedBoost` *satisfy*

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\Big[\max\Big\{\Big|Q^{+}_{\gamma/2}(x_i) - y_i\Big|, \Big|Q^{-}_{\gamma/2}(x_i) - y_i\Big|\Big\} > \eta/2\Big] \leq \prod_{t=1}^{T} e^{\gamma\alpha_t}\sum_{i=1}^{m} P_t(i)e^{-\alpha_t\theta_i^{(t)}}.$$

We note that, in the special case of binary classification, `MedBoost` is closely related to the well-known AdaBoost algorithm (Freund and Schapire, 1997), and the above results correspond to a standard margin-based analysis of Schapire et al. (1998). For our purposes, we will need the following immediate corollary of this, which follows from plugging in the values of α_t and using the weak learning assumption, which implies $\sum_{i=1}^{m} P_t(i)\mathbb{I}[\theta_i^{(t)} = 1] \geq \frac{1}{2} + \gamma$ for all t.

Corollary 6 *For $T = \Theta\Big(\frac{1}{\gamma^2}\ln(m)\Big)$, every $i \in \{1, \ldots, m\}$ has*

$$\max\Big\{\Big|Q^{+}_{\gamma/2}(x_i) - y_i\Big|, \Big|Q^{-}_{\gamma/2}(x_i) - y_i\Big|\Big\} \leq \eta/2.$$

6. The Sample Complexity of Learning Real-Valued Functions

This section reveals our intention in choosing this notion of weak hypothesis, rather than using, say, an ε-good strong learner under absolute loss. In addition to being a strong enough notion for boosting to work, we show here that it is also a weak enough notion for the sample complexity of weak learning to be of reasonable size: namely, a size quantified by the fat-shattering dimension. This result is also relevant to an open question posed by Simon (1997), who proved a lower bound for the sample complexity of finding an (η, γ)-weak hypothesis, expressed in terms of a related complexity measure, and asked whether a related upper bound might also hold. We establish a general upper bound here, witnessing the same dependence on the parameters η and γ as observed in Simon's lower bound (up to a log factor) aside from a difference in the key complexity measure appearing in the bounds.

Define $\rho_\eta(f, g) = P_{2m}(x : |f(x) - g(x)| > \eta)$, where P_{2m} is the empirical measure induced by $X_1, \ldots, X_{2m}$ iid P-distributed random variables (the m data points and m ghost points). Define $N_\eta(\beta)$ as the β-covering numbers of $\mathcal{F}$ under the ρ_η pseudo-metric.

Theorem 7 *Fix any $\eta, \beta \in (0, 1)$, $\alpha \in [0, 1)$, $m \in \mathbb{N}$ and a target function f^*. For $X_1, \ldots, X_m$ iid P-distributed, with probability at least $1 - \mathbb{E}\big[N_{\eta(1-\alpha)/2}(\beta/8)\big]\, 2e^{-m\beta/96}$, every $f \in \mathcal{F}$ with $\max_{1\leq i\leq m}|f(X_i) - f^*(X_i)| \leq \alpha\eta$ satisfies $P(x : |f(x) - f^*(x)| > \eta) \leq \beta$.*

Proof This proof roughly follows the usual symmetrization argument for uniform convergence Vapnik and Červonenkis (1971); Haussler (1992), with a few important modifications to account for this (η, β)-based criterion. If $\mathbb{E}\big[N_{\eta(1-\alpha)/2}(\beta/8)\big]$ is infinite, then the result is trivial, so let us suppose it is finite for the remainder of the proof. Similarly, if $m < 8/\beta$, then $2e^{-m\beta/96} > 1$ and hence the claim trivially holds, so let us suppose $m \geq 8/\beta$ for the remainder of the proof. Without loss of generality, suppose $f^*(x) = 0$ everywhere and every $f \in \mathcal{F}$ is non-negative (otherwise subtract f^* from every $f \in \mathcal{F}$ and redefine $\mathcal{F}$ as the absolute values of the differences; note that this transformation does not increase the value of $N_{\eta(1-\alpha)/2}(\beta/8)$ since applying this transformation to the original $N_{\eta(1-\alpha)/2}(\beta/8)$ functions remains a cover).

Let $X_1, \ldots, X_{2m}$ be iid P-distributed. Denote by P_m the empirical measure induced by $X_1, \ldots, X_m$, and by P'_m the empirical measure induced by $X_{m+1}, \ldots, X_{2m}$. We have

$$\begin{aligned}&\mathbb{P}\big(\exists f \in \mathcal{F} : P'_m(x : f(x) > \eta) > \beta/2 \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1\big) \\ &\geq \mathbb{P}\left(\exists f \in \mathcal{F} : P(x : f(x) > \eta) > \beta \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1 \text{ and } P'_m(x : f(x) > \eta) > \beta/2\right).\end{aligned}$$

Denote by A_m the event that there exists $f \in \mathcal{F}$ satisfying $P(x : f(x) > \eta) > \beta$ and $P_m(x : f(x) \leq \alpha\eta) = 1$, and on this event let $\tilde{f}$ denote such an $f \in \mathcal{F}$ (chosen solely based on $X_1, \ldots, X_m$); when A_m fails to hold, take $\tilde{f}$ to be some arbitrary fixed element of $\mathcal{F}$. Then the expression on the right hand side above is at least as large as

$$\mathbb{P}\left(A_m \text{ and } P'_m(x : \tilde{f}(x) > \eta) > \beta/2\right),$$

and noting that the event A_m is independent of $X_{m+1}, \ldots, X_{2m}$, this equals

$$\mathbb{E}\left[\mathbb{I}_{A_m} \cdot \mathbb{P}\left(P'_m(x : \tilde{f}(x) > \eta) > \beta/2 \middle| X_1, \ldots, X_m\right)\right]. \tag{1}$$

Then note that for any $f \in \mathcal{F}$ with $P(x : f(x) > \eta) > \beta$, a Chernoff bound implies

$$\begin{aligned}&\mathbb{P}\Big(P'_m(x : f(x) > \eta) > \beta/2\Big) \\ &= 1 - \mathbb{P}\Big(P'_m(x : f(x) > \eta) \leq \beta/2\Big) \geq 1 - \exp\{-m\beta/8\} \geq \frac{1}{2},\end{aligned}$$

where we have used the assumption that $m \geq \frac{8}{\beta}$ here. In particular, this implies that the expression in (1) is no smaller than $\frac{1}{2}\mathbb{P}(A_m)$. Altogether, we have established that

$$\begin{aligned}&\mathbb{P}(\exists f \in \mathcal{F} : P(x : f(x) > \eta) > \beta \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1) \\ &\leq 2\mathbb{P}\big(\exists f \in \mathcal{F} : P'_m(x : f(x) > \eta) > \beta/2 \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1\big).\end{aligned} \tag{2}$$

Now let $\sigma(1), \ldots, \sigma(m)$ be independent random variables (also independent of the data), with $\sigma(i) \sim \text{Uniform}(\{i, m+i\})$, and denote $\sigma(m+i)$ as the sole element of $\{i, m+i\} \setminus \{\sigma(i)\}$ for each $i \leq m$. Also denote by $P_{m,\sigma}$ the empirical measure induced by $X_{\sigma(1)}, \ldots, X_{\sigma(m)}$, and by $P'_{m,\sigma}$ the empirical measure induced by $X_{\sigma(m+1)}, \ldots, X_{\sigma(2m)}$. By exchangeability of $(X_1, \ldots, X_{2m})$, the probability on the right hand side of (2) is equal to

$$\mathbb{P}\big(\exists f \in \mathcal{F} : P'_{m,\sigma}(x : f(x) > \eta) > \beta/2 \text{ and } P_{m,\sigma}(x : f(x) \leq \alpha\eta) = 1\big).$$

Now let $\hat{\mathcal{F}} \subseteq \mathcal{F}$ be a minimal subset of $\mathcal{F}$ such that $\max\limits_{f \in \mathcal{F}} \min\limits_{\hat{f} \in \hat{\mathcal{F}}} \rho_{\eta(1-\alpha)/2}(\hat{f}, f) \leq \beta/8$. The size of $\hat{\mathcal{F}}$ is at most $N_{\eta(1-\alpha)/2}(\beta/8)$, which is finite almost surely (since we have assumed above that its expectation is finite). Then note that (denoting by $X_{[2m]} = (X_1, \ldots, X_{2m})$) the above expression is at most

$$\begin{aligned}&\mathbb{P}\Big(\exists f \in \hat{\mathcal{F}} : P'_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) > (3/8)\beta \text{ and } P_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) \leq \beta/8\Big) \\ &\leq \mathbb{E}\Bigg[N_{\eta(1-\alpha)/2}(\beta/8) \max_{f \in \hat{\mathcal{F}}} \mathbb{P}\big(P'_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) > (3/8)\beta \\ &\qquad\qquad \text{and } P_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) \leq \beta/8 \big| X_{[2m]}\big)\Bigg].\end{aligned} \tag{3}$$

Then note that for any $f \in \mathcal{F}$, we have almost surely

$$\begin{aligned}
&\mathbb{P}\big(P'_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) > (3/8)\beta \text{ and } P_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) \le \beta/8 \big| X_{[2m]}\big) \\
&\le \mathbb{P}\big(P_{2m}(x : f(x) > \eta(1+\alpha)/2) > (3/16)\beta \text{ and } P_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) \le \beta/8 \big| X_{[2m]}\big) \\
&\le \exp\{-m\beta/96\}\,,
\end{aligned}$$

where the last inequality is by a Chernoff bound, which (as noted by Hoeffding (1963)) remains valid even when sampling without replacement. Together with (2) and (3), we have that

$$\begin{aligned}
&\mathbb{P}(\exists f \in \mathcal{F} : P(x : f(x) > \eta) > \beta \text{ and } P_m(x : f(x) \le \alpha\eta) = 1) \\
&\le 2\mathbb{E}\big[N_{\eta(1-\alpha)/2}(\beta/8)\big]\, e^{-m\beta/96}.
\end{aligned}$$

■

The following lemma is also new. Together with Theorem 7, it enables us to express the sample complexity as a simple function of the fat-shattering dimension.

Lemma 8 *There exist universal numerical constants $c, c' \in (0,\infty)$ such that $\forall \eta, \beta \in (0,1)$,*

$$N_\eta(\beta) \le \left(\frac{2}{\eta\beta}\right)^{cd(c'\eta\beta)},$$

where $d(\cdot)$ is the fat-shattering dimension.

Proof Mendelson and Vershynin (2003, Theorem 1) establishes that the $\eta\beta$-covering number of $\mathcal{F}$ under the $L_2(P_{2m})$ pseudo-metric is at most

$$\left(\frac{2}{\eta\beta}\right)^{cd(c'\eta\beta)} \tag{4}$$

for some universal numerical constants $c, c' \in (0,\infty)$. Then note that for any $f, g \in \mathcal{F}$, Markov's and Jensen's inequalities imply $\rho_\eta(f,g) \le \frac{1}{\eta}\|f-g\|_{L_1(P_{2m})} \le \frac{1}{\eta}\|f-g\|_{L_2(P_{2m})}$. Thus, any $\eta\beta$-cover of $\mathcal{F}$ under $L_2(P_{2m})$ is also a β-cover of $\mathcal{F}$ under ρ_η, and therefore (4) is also a bound on $N_\eta(\beta)$. ■

Combining the above two results yields the following theorem.

Theorem 9 *For some universal numerical constants $c_1, c_2, c_3 \in (0,\infty)$, for any $\eta, \delta, \beta \in (0,1)$ and $\alpha \in [0,1)$, letting $X_1, \dots, X_m$ be iid P-distributed, where*

$$m = \left\lceil \frac{c_1}{\beta}\left(d(c_2\eta\beta(1-\alpha))\ln\left(\frac{c_3}{\eta\beta(1-\alpha)}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right\rceil,$$

with probability at least $1-\delta$, every $f \in \mathcal{F}$ with $\max_{i\in[m]}|f(X_i) - f^*(X_i)| \le \alpha\eta$ satisfies $P(x : |f(x) - f^*(x)| > \eta) \le \beta$.

Proof The result follows immediately from combining Theorem 7 and Lemma 8. ∎

In particular, Theorem 4 follows immediately from this result by taking $\beta = 1/2 - \gamma$ and $\alpha = \gamma/2$.

To discuss tightness of Theorem 9, we note that Simon (1997) proved a sample complexity lower bound for the same criterion of

$$\Omega\left(\frac{d'(c\eta)}{\beta} + \frac{1}{\beta}\log\frac{1}{\delta}\right),$$

where $d'(\cdot)$ is a quantity somewhat smaller than the fat-shattering dimension, essentially representing a fat Natarajan dimension. Thus, aside from the differences in the complexity measure (and a logarithmic factor), we establish an upper bound of a similar form to Simon's lower bound.

7. From Boosting to Compression

Generally, our strategy for converting the boosting algorithm `MedBoost` into a sample compression scheme of smaller size follows a strategy of Moran and Yehudayoff for binary classification, based on arguing that because the ensemble makes its predictions with a *margin* (corresponding to the results on *quantiles* in Corollary 6), it is possible to recover the same proximity guarantees for the predictions while using only a smaller *subset* of the functions from the original ensemble. Specifically, we use the following general *sparsification* strategy.

For $\alpha_1, \ldots, \alpha_T \in [0,1]$ with $\sum_{t=1}^{T} \alpha_t = 1$, denote by $\mathrm{Cat}(\alpha_1, \ldots, \alpha_T)$ the *categorical distribution*: i.e., the discrete probability distribution on $\{1, \ldots, T\}$ with probability mass α_t on t.

Algorithm 2: `Sparsify`$(\{(x_i, y_i)\}_{i\in[m]}, \gamma, \eta, T, n)$

Run `MedBoost`$(\{(x_i, y_i)\}_{i\in[m]}, T, \gamma, \eta)$
Let $h_1, \ldots, h_T$ and $\alpha_1, \ldots, \alpha_T$ be its return values
Denote $\alpha'_t = \alpha_t / \sum_{t'=1}^{T} \alpha_{t'}$ for each $t \in [T]$
repeat
    Sample $(J_1, \ldots, J_n) \sim \mathrm{Cat}(\alpha'_1, \ldots, \alpha'_T)^n$
    Let $F = \{h_{J_1}, \ldots, h_{J_n}\}$
until $\max_{1\le i\le m} |\{f \in F : |f(x_i) - y_i| > \eta\}| < n/2$
Return F

For any values $a_1, \ldots, a_n$, denote the (unweighted) median

$$\mathrm{Med}(a_1, \ldots, a_n) = \mathrm{Median}(a_1, \ldots, a_n; 1, \ldots, 1).$$

Our intention in dicussing the above algorithm is to argue that, for a sufficiently large choice of n, the above procedure returns a set $\{f_1, \ldots, f_n\}$ such that

$$\forall i \in [m], |\,\mathrm{Med}(f_1(x_i), \ldots, f_n(x_i)) - y_i| \le \eta.$$

We analyze this strategy separately for binary classification and real-valued functions, since the argument in the binary case is much simpler (and demonstrates more directly the connection to the original argument of Moran and Yehudayoff), and also because we arrive at a tighter result for binary functions than for real-valued functions.

7.1. Binary Classification

We begin with the simple observation about binary classification (i.e., where the functions in $\mathcal{F}$ all map into $\{0,1\}$). The technique here is quite simple, and follows a similar line of reasoning to the original argument of Moran and Yehudayoff. The argument for real-valued functions below will diverge from this argument in important ways, requiring several non-trivial new techniques in the proof, though the high-level outline of the argument remains the same.

The compression function is essentially the one introduced by Moran and Yehudayoff, except applied to the classifiers produced by the above `Sparsify` procedure, rather than a set of functions selected by a minimax distribution over all classifiers produced by $O(d)$ samples each. The weak hypotheses in `MedBoost` for binary classification can be obtained using samples of size $O(d)$. Thus, if the `Sparsify` procedure is successful in finding n such classifiers whose median predictions are within η of the target y_i values for all i, then we may encode these n classifiers as a compression set, consisting of the set of $k = O(nd)$ samples used to train these classifiers, together with $k \log k$ extra bits to encode the order of the samples.[2] To obtain Theorem 1, it then suffices to argue that $n = \Theta(d^*)$ is a sufficient value. The proof follows.

Proof [Proof of Theorem 1] Recall that d^* bounds the VC dimension of the class of sets $\{\{h_t : t \le T, h_t(x_i) = 1\} : 1 \le i \le m\}$. Thus for the iid samples $h_{J_1}, \ldots, h_{J_n}$ obtained in `Sparsify`, for $n = 64(2309 + 16d^*) > \frac{2304+16d^*+\log(2)}{(1/8)^2}$, by the VC uniform convergence inequality of Vapnik and Červonenkis (1971), with probability at least $1/2$ we get that

$$\max_{1\le i\le m} \left| \left(\frac{1}{n} \sum_{j=1}^{n} h_{J_j}(x_i) \right) - \left(\sum_{t=1}^{T} \alpha'_t h_t(x_i) \right) \right| < 1/8.$$

In particular, if we choose $\eta = 1/2$, $\gamma = 1/4$, and $T = \Theta(\log(m))$ appropriately, then Corollary 6 implies that every $Q^+_{\gamma/2}(x_i)$ and $Q^-_{\gamma/2}(x_i)$ must both be equal y_i (using the fact that if two binary-valued quantities have distance strictly less than 1 then they must be equal). This implies every $y_i = \mathbb{I}\left[\sum_{t=1}^{T} \alpha'_t h_t(x_i) \ge 1/2\right]$ and $\left|\frac{1}{2} - \sum_{t=1}^{T} \alpha'_t h_t(x_i)\right| \ge 1/8$ so that the above event would imply every $y_i = \mathbb{I}\left[\frac{1}{n}\sum_{j=1}^{n} h_{J_j}(x_i) \ge 1/2\right] = \mathrm{Med}(h_{J_1}(x_i), \ldots, h_{J_n}(x_i))$. Note that the `Sparsify` algorithm need only try this sampling $\log_2(1/\delta)$ times to find such a set of n functions. Combined with the description above (from Moran and Yehudayoff, 2016) of how to encode this collection of h_{J_i} functions as a sample compression set plus side information, this completes the construction of the sample compression scheme. ■

7.2. Real-Valued Functions

Next we turn to the general case of real-valued functions (where the functions in $\mathcal{F}$ may generally map into $[0,1]$). We have the following result, which says that the `Sparsify` procedure can reduce the ensemble of functions from one with $T = O(\log(m)/\gamma^2)$ functions in it, down to one with a number of functions *independent of* m.

2. In fact, $k \log n$ bits would suffice if the weak learner is permutation-invariant in its data set.

Theorem 10 *Choosing*

$$n = \Theta\bigg(\frac{1}{\gamma^2} d^*(c\eta) \log^2(d^*(c\eta)/\eta)\bigg)$$

suffices for the `Sparsify` *procedure to return* $\{f_1, \ldots, f_n\}$ *with*

$$\max_{1\le i\le m} |\operatorname{Med}(f_1(x_i), \ldots, f_n(x_i)) - y_i| \le \eta.$$

Proof Recall from Corollary 6 that `MedBoost` returns functions $h_1, \ldots, h_T \in \mathcal{F}$ and $\alpha_1, \ldots, \alpha_T \ge 0$ such that $\forall i \in \{1, \ldots, m\}$,

$$\max\Big\{\Big|Q^+_{\gamma/2}(x_i) - y_i\Big|, \Big|Q^-_{\gamma/2}(x_i) - y_i\Big|\Big\} \le \eta/2, \tag{5}$$

where $\{(x_i, y_i)\}_{i=1}^m$ is the training data set.

We use this property to sparsify $\{h_1, \ldots, h_T\}$ from $T = O(\log(m)/\gamma^2)$ down to k elements, where k will depend on η, γ, and the dual fat-shattering dimension of $\mathcal{F}$ (actually, just of $H = \{h_1, \ldots, h_T\} \subseteq \mathcal{F}$) — but **not** sample size m.

First, we define the offset class $\bar{H} = H - f^* = \{\hbar_t = h_t - f^* : t \in [T]\}$. By appropriately choosing the offset $r \in \mathbb{R}^m$, we see that $\bar{H}$ has the same fat-shattering dimension at all scales as H.

Letting $\alpha'_j = \alpha_j / \sum_{t=1}^T \alpha_t$ for each $j \le T$, we will sample k hypotheses $\Big\{\tilde{h}_1, \ldots, \tilde{h}_k\Big\} =: \tilde{H} \subseteq H$ with each $\tilde{h}_i = h_{J_i}$, where $(J_1, \ldots, J_k) \sim \operatorname{Cat}(\alpha'_1, \ldots, \alpha'_T)^k$ as in `Sparsify`. Each $\tilde{h}_i$ will have a corresponding $\hbar_i = \tilde{h}_i - f*$. Define the functions $\hat{h}(x) = \operatorname{Med}(\tilde{h}_1(x), \ldots, \tilde{h}_k(x))$ and $\hat{\hbar}(x) = \operatorname{Med}(\hbar_1(x), \ldots, \hbar_k(x))$. We claim that for any fixed $i \in [m]$, with high probability

$$|\hat{h}(x_i) - f^*(x_i)| = |\hbar(x_i)| \le \eta/2. \tag{6}$$

Indeed, since for all $\boldsymbol{\xi} \in \mathbb{R}^m$ and $a \in \mathbb{R}$, we have $\operatorname{Med}(\boldsymbol{\xi}) + a = \operatorname{Med}(\boldsymbol{\xi} + a)$, it follows that (5) may be rewritten as

$$\max\big\{\big|Q^+_\gamma(\hbar_1(x), \ldots, \hbar_T(x); \alpha_1, \ldots, \alpha_T)\big|, \big|Q^-_\gamma(\hbar_1(x), \ldots, \hbar_T(x); \alpha_1, \ldots, \alpha_T)\big|\big\} \le \eta/2. \tag{7}$$

Partition the indices $[T]$ into the disjoint sets

$$\begin{aligned}
L(x) &= \big\{j \in [T] : \hbar_j(x) < Q^-_\gamma(\hbar_1(x), \ldots, \hbar_T(x); \alpha_1, \ldots, \alpha_T)\big\}, \\
M(x) &= \big\{j \in [T] : Q^-_\gamma(\hbar_1(x), ..., \hbar_T(x); \alpha_1, ..., \alpha_T) \le \hbar_j(x) \le Q^+_\gamma(\hbar_1(x), ..., \hbar_T(x); \alpha_1, ..., \alpha_T)\big\}, \\
R(x) &= \big\{j \in [T] : \hbar_j(x) > Q^+_\gamma(\hbar_1(x), \ldots, \hbar_T(x); \alpha_1, \ldots, \alpha_T)\big\}.
\end{aligned}$$

Then the only way (6) can fail is if half or more indices $J_1, \ldots, J_k$ sampled fall into $R(x_i)$ — or if half or more fall into $L(x_i)$. Since the sampling distribution puts mass less than $1/2 - \gamma$ on each of $R(x_i)$ and $L(x_i)$, Chernoff's bound puts an upper estimate of $\exp(-2k\gamma^2)$ on either event. Hence,

$$\mathbb{P}\left(|\hat{h}(x_i) - f^*(x_i)| > \eta/2\right) = \mathbb{P}\left(|\hat{\hbar}(x_i)| > \eta/2\right) \le 2\exp(-2k\gamma^2). \tag{8}$$

Next, our goal is to ensure that with high probability, (6) holds simultaneously for all $i \in [m]$. Define the map $\boldsymbol{\xi} : [m] \to \mathbb{R}^k$ by $\boldsymbol{\xi}(i) = (\hbar_1(x_i), \ldots, \hbar_k(x_i))$. Let $G \subseteq [m]$ be a minimal subset of $[m]$ such that

$$\max_{i \in [m]} \min_{j \in G} \|\boldsymbol{\xi}(i) - \boldsymbol{\xi}(j)\|_\infty \le \eta/2.$$

This is just a minimal ℓ_∞ covering of $[m]$. Then

$$\begin{aligned}
&\mathbb{P}\left(\exists i \in [m] : |\operatorname{Med}(\boldsymbol{\xi}(i))| > \eta\right) \le \\
&\sum_{j \in G} \mathbb{P}\left(\exists i : |\operatorname{Med}(\boldsymbol{\xi}(i))| > \eta, \|\boldsymbol{\xi}(i) - \boldsymbol{\xi}(j)\|_\infty \le \eta/2\right) \le \\
&\sum_{j \in G} \mathbb{P}\left(|\operatorname{Med}(\boldsymbol{\xi}(j))| > \eta/2\right) \le 2N_\infty([m], \eta/2) \exp(-2k\gamma^2),
\end{aligned}$$

where $N_\infty([m], \eta/2)$ is the $\eta/2$-covering number (under ℓ_∞) of $[m]$, and we used the fact that

$$|\operatorname{Med}(\boldsymbol{\xi}(i)) - \operatorname{Med}(\boldsymbol{\xi}(j))| \le \|\boldsymbol{\xi}(i) - \boldsymbol{\xi}(j)\|_\infty .$$

Finally, to bound $N_\infty([m], \eta/2)$, note that $\boldsymbol{\xi}$ embeds $[m]$ into the dual class $\mathcal{F}^*$. Thus, we may apply the bound in (Rudelson and Vershynin, 2006, Display (1.4)):

$$\log N_\infty([m], \eta/2) \le C d^*(c\eta) \log^2(k/\eta),$$

where C, c are universal constants and $d^*(\cdot)$ is the dual fat-shattering dimension of $\mathcal{F}$. It now only remains to choose a k that makes $\exp\left(C d^*(c\eta) \log^2(k/\eta) - 2k\gamma^2\right)$ as small as desired. ■

To establish Theorem 2, we use the weak learner from above, with the booster `MedBoost` from Kégl, and then apply the `Sparsify` procedure. Combining the corresponding theorems, together with the same technique for converting to a compression scheme discussed above for classification (i.e., encoding the functions with the set of training examples they were obtained from, plus extra bits to record the order and which examples which weak hypothesis was obtained by training on), this immediately yields the result claimed in Theorem 2, which represents our main new result for sample compression of general families of real-valued functions.

Acknowledgments

This research was supported by the Lynn and William Frankel Center for Computer Science at Ben-Gurion University and by the Israel Science Foundation.
We thank Shay Moran and Roi Livni for insightful conversations.

References

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997. URL citeseer.ist.psu.edu/alon97scalesensitive.html.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999. ISBN 0-521-57353-X. doi: 10.1017/CBO9780511624216. URL http://dx.doi.org/10.1017/CBO9780511624216.

Martin Anthony, Peter L. Bartlett, Yuval Ishai, and John Shawe-Taylor. Valid generalisation from approximate interpolation. *Combinatorics, Probability & Computing*, 5:191–214, 1996. doi: 10.1017/S096354830000198X. URL https://doi.org/10.1017/S096354830000198X.

Patrice Assouad. Densité et dimension. *Ann. Inst. Fourier (Grenoble)*, 33(3):233–282, 1983. ISSN 0373-0956. URL http://www.numdam.org/item?id=AIF_1983__33_3_233_0.

Ran Avnimelech and Nathan Intrator. Boosting regression estimators. *Neural computation*, 11(2):499–520, 1999.

Shai Ben-David and Ami Litman. Combinatorial variability of vapnik-chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, 86(1):3–25, 1998.

Alberto Bertoni, Paola Campadelli, and M Parodi. A boosting algorithm for regression. In *International Conference on Artificial Neural Networks*, pages 343–348. Springer, 1997.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. ISSN 0004-5411.

Artem Chernikov and Pierre Simon. Externally definable sets and dependent pairs. *Israel J. Math.*, 194(1):409–425, 2013. ISSN 0021-2172. URL https://doi.org/10.1007/s11856-012-0061-9.

Ofir David, Shay Moran, and Amir Yehudayoff. Supervised learning through the lens of compression. In *Advances in Neural Information Processing Systems*, pages 2784–2792, 2016.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996. ISBN 0-387-94618-7.

Harris Drucker. Improving regressors using boosting techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 107–115, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1-55860-486-3. URL http://dl.acm.org/citation.cfm?id=645526.657132.

Richard M. Dudley. A course on empirical processes. In *École d'été de probabilités de Saint-Flour, XII—1982*, volume 1097 of *Lecture Notes in Math.*, pages 1–142. Springer, Berlin, 1984.

Nigel Duffy and David Helmbold. Boosting methods for regression. *Machine Learning*, 47:153–200, 2002. ISSN 0885-6125.

Mary Flahive and Bella Bose. Balancing cyclic r-ary gray codes. *the electronic journal of combinatorics*, 14(1):R31, 2007.

Sally Floyd. Space-bounded learning and the vapnik-chervonenkis dimension. In *Proceedings of the second annual workshop on Computational learning theory*, pages 349–364. Morgan Kaufmann Publishers Inc., 1989.

Sally Floyd and Manfred K. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Yoav Freund and Robert E Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the ninth annual conference on Computational learning theory*, pages 325–332. ACM, 1996.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997. ISSN 0022-0000. doi: http://dx.doi.org/10.1006/jcss.1997.1504.

Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games Econom. Behav.*, 29(1-2):79–103, 1999. ISSN 0899-8256. URL https://doi.org/10.1006/game.1999.0738. Learning in games: a symposium in honor of David Blackwell.

Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 2001. ISSN 0090-5364. URL https://doi.org/10.1214/aos/1013203451.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 370–378, 2014.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Trans. Information Theory*, 63(8):4838–4849, 2017a. doi: 10.1109/TIT.2017.2713820. URL https://doi.org/10.1109/TIT.2017.2713820.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, Aug 2017b. ISSN 0018-9448. doi: 10.1109/TIT.2017.2713820.

Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. PAC-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59 (1-2):55–76, 2005.

S. Hanneke, A. Kontorovich, and M. Sadigurschi. Agnostic sample compression for linear regression. *In Submission*, 2018. URL http://www.stevehanneke.com/docs/2018/HKS-linear.pdf.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992. doi: 10.1016/0890-5401(92)90010-D. URL http://dx.doi.org/10.1016/0890-5401(92)90010-D.

David Haussler, Michael Kearns, Nick Littlestone, and Manfred K Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*, 95(2):129–161, 1991.

David Helmbold, Robert Sloan, and Manfred K Warmuth. Learning integer lattices. *SIAM Journal on Computing*, 21(2):240–266, 1992.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963.

Daniel M. Kane, Roi Livni, Shay Moran, and Amir Yehudayoff. On communication complexity of classification problems. *CoRR*, abs/1711.05893, 2017. URL http://arxiv.org/abs/1711.05893.

Grigoris Karakoulas and John Shawe-Taylor. Towards a strategy for boosting regressors. In Alexander J. Smola, Peter L. Bartlett, and Schölkopf, editors, *Advances in Large Margin Classifiers*, Advances in Neural Information Processing Systems, pages 43–54. MIT Press, Cambridge, MA, USA, 2000. ISBN 0-262-19448-1.

Balázs Kégl. Robust regression by boosting the median. In *Learning Theory and Kernel Machines*, pages 258–272. Springer, 2003.

Dima Kuzmin and Manfred K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007. URL http://dl.acm.org/citation.cfm?id=1314566.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability. Technical report, Department of Computer and Information Sciences, Santa Cruz, CA, Ju, 1986.

Roi Livni and Pierre Simon. Honest compressions and their application to compression schemes. In *Conference on Learning Theory*, pages 77–92, 2013.

Philip M. Long. Efficient algorithms for learning functions with bounded variation. *Inf. Comput.*, 188(1):99–115, 2004. doi: 10.1016/S0890-5401(03)00164-0. URL https://doi.org/10.1016/S0890-5401(03)00164-0.

Shie Mannor and Ron Meir. On the existence of linear weak learners and applications to boosting. *Machine Learning*, 48(1-3):219–251, 2002. doi: 10.1023/A:1013959922467. URL https://doi.org/10.1023/A:1013959922467.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 512–518, Cambridge, MA, USA, 1999. MIT Press. URL http://dl.acm.org/citation.cfm?id=3009657.3009730.

S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Invent. Math.*, 152(1):37–55, 2003. ISSN 0020-9910. doi: 10.1007/s00222-002-0266-3. URL http://dx.doi.org/10.1007/s00222-002-0266-3.

João Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. Ensemble approaches for regression: A survey. *ACM Comput. Surv.*, 45(1):10:1–10:40, December 2012. ISSN 0360-0300. doi: 10.1145/2379776.2379786. URL http://doi.acm.org/10.1145/2379776.2379786.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3):21:1–21:10, 2016. doi: 10.1145/2890490. URL http://doi.acm.org/10.1145/2890490.

Shay Moran, Amir Shpilka, Avi Wigderson, and Amir Yehudayoff. Teaching and compressing for low vc-dimension. In *A Journey Through Discrete Mathematics*, pages 633–656. Springer, 2017.

Richard Nock and Frank Nielsen. A real generalization of discrete adaboost. *Artificial Intelligence*, 171(1):25 – 41, 2007. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2006.10.014. URL http://www.sciencedirect.com/science/article/pii/S0004370206001111.

Leonard Pitt and Leslie G Valiant. Computational limitations on learning from examples. *Journal of the ACM (JACM)*, 35(4):965–984, 1988.

David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

Benjamin I. P. Rubinstein and J. Hyam Rubinstein. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13:1221–1261, 2012. URL http://dl.acm.org/citation.cfm?id=2343686.

Benjamin I. P. Rubinstein, Peter L. Bartlett, and J. Hyam Rubinstein. Shifting: One-inclusion mistake bounds and sample compression. *J. Comput. Syst. Sci.*, 75(1):37–59, 2009. doi: 10.1016/j.jcss.2008.07.005. URL https://doi.org/10.1016/j.jcss.2008.07.005.

M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Ann. of Math. (2)*, 164(2):603–648, 2006. ISSN 0003-486X. URL https://doi.org/10.4007/annals.2006.164.603.

Robert E. Schapire and Yoav Freund. *Boosting*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2012. ISBN 978-0-262-01718-3. Foundations and algorithms.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, 26(5):1651–1686, 1998. ISSN 0090-5364. doi: 10.1214/aos/1024691352. URL http://dx.doi.org/10.1214/aos/1024691352.

Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. *SIAM J. Comput.*, 26(3):751–763, 1997. doi: 10.1137/S0097539793259185. URL https://doi.org/10.1137/S0097539793259185.

V. N. Vapnik and A. Ja. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279, 1971. ISSN 0040-361x.

V. N. Vapnik and A. Ya. Chervonenkis. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya*. Izdat. "Nauka", Moscow, 1974.

Appendix A. Sample compression for BV functions

The function class $\mathrm{BV}(v)$ consists of all $f : [0,1] \to \mathbb{R}$ for which

$$V(f) := \sup_{n\in\mathbb{N}} \sup_{0=x_0<x_1<\ldots<x_n=1} \sum_{i=1}^{n-1} |f(x_{i+1}) - f(x_i)| \le v.$$

It is known (Anthony and Bartlett, 1999, Theorem 11.12) that $d_{\mathrm{BV}(v)}(t) = 1 + \lfloor v/(2t) \rfloor$. In Theorem 12 below, we show that the dual class has $d^*_{\mathrm{BV}(v)}(t) = \Theta\left(\log(v/t)\right)$. Long (2004) presented an efficient, proper, consistent learner for the class $\mathcal{F} = \mathrm{BV}(1)$ with range restricted to $[0,1]$, with sample complexity $m_{\mathcal{F}}(\varepsilon, \delta) = O(\frac{1}{\varepsilon} \log \frac{1}{\delta})$. Combined with Theorem 2, this yields

Corollary 11 *Let $\mathcal{F} = \mathrm{BV}(1) \cap [0,1]^{[0,1]}$ be the class $f : [0,1] \to [0,1]$ with $V(f) \le 1$. Then the proper, consistent learner $\mathcal{L}$ of Long (2004), with target generalization error ε, admits a sample compression scheme of size $O(k \log k)$, where*

$$k = O\left(\frac{1}{\varepsilon} \log^2 \frac{1}{\varepsilon} \cdot \log\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right)\right).$$

The compression set is computable in expected runtime

$$O\left(n \frac{1}{\varepsilon^{3.38}} \log^{3.38} \frac{1}{\varepsilon} \left(\log n + \log \frac{1}{\varepsilon} \log\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right)\right)\right).$$

The remainder of this section is devoted to proving

Theorem 12 *For $\mathcal{F} = \mathrm{BV}(v)$ and $t < v$, we have $d^*_{\mathcal{F}}(t) = \Theta\left(\log(v/t)\right)$.*

First, we define some preliminary notions:

Definition 13 *For a binary $m \times n$ matrix M, define*

$$\begin{aligned} V(M,i) &:= \sum_{j=1}^{m} \mathbb{I}[M_{j,i} \ne M_{j+1,i}], \\ G(M) &:= \sum_{i=1}^{n} V(M,i), \\ V(M) &:= \max_{i\in[n]} V(M,i). \end{aligned}$$

Lemma 14 *Let M be a binary $2^n \times n$ matrix. If for each $b \in \{0,1\}^n$ there is a row j in M equal to b, then*

$$V(M) \geq \frac{2^n}{n}.$$

In particular, for at least one row i, we have $V(M,i) \geq 2^n/n$.

Proof Let M be a $2^n \times n$ binary such that for each $b \in \{0,1\}^n$ there is a row j in M equal to b. Given M's dimensions, every $b \in \{0,1\}^n$ appears exactly in one row of M, and hence the minimal Hamming distance between two rows is 1. Summing over the $2^n - 1$ adjacent row pairs, we have

$$G(M) = \sum_{i=1}^{n} V(M,i) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{I}[M_{j,i} \neq M_{j+1,i}] \geq 2^n - 1,$$

which averages to

$$\frac{1}{n}\sum_{i=1}^{n} V(M,i) = \frac{G(M)}{n} \geq \frac{2^n - 1}{n}.$$

By the pigeon-hole principle, there must be a row $j \in [n]$ for which $V(M,i) \geq \frac{2^n-1}{n}$, which implies $V(M) \geq \frac{2^n-1}{n}$. ■

We split the proof of Theorem 12 into two estimates:

Lemma 15 *For $\mathcal{F} = \mathrm{BV}(v)$ and $t < v$, $d^*_{\mathcal{F}}(t) \leq 2\log_2(v/t)$.*

Lemma 16 *For $\mathcal{F} = \mathrm{BV}(v)$ and $4t < v$, $d^*_{\mathcal{F}}(t) \geq \lfloor \log_2(v/t) \rfloor$.*

Proof [Proof of Lemma 15] Let $\{f_1, \ldots, f_n\} \subset \mathcal{F}$ be a set of functions that are t-shattered by $\mathcal{F}^*$. In other words, there is an $r \in \mathbb{R}^n$ such that for each $b \in \{0,1\}^n$ there is an $x_b \in \mathcal{F}^*$ such that

$$\forall i \in [n], x_b(f_i) \begin{cases} \geq r_i + t, & b_i = 1 \\ \leq r_i - t, & b_i = 0 \end{cases}.$$

Let us order the x_bs by magnitude $x_1 < x_2 < \ldots < x_{2^n}$, denoting this sequence by $(x_i)_{i=1}^{2^n}$. Let $M \in \{0,1\}^{2^n \times n}$ be a matrix whose ith row is b_j, the latter ordered arbitrarily.

By Lemma 14, there is $i \in [n]$ s.t.

$$\sum_{j=1}^{2^n} \mathbb{I}[M(j,i) \neq M(j+1,i)] \geq \frac{2^n}{n}.$$

Note that if $M(j,i) \neq M(j+1,i)$ shattering implies that

$$x_j(f_i) \geq r_i + t \text{ and } x_{j+1}(f_i) \leq r_i - t$$

or

$$x_j(f_i) \leq r_i - t \text{ and } x_{j+1}(f_i) \geq r_i + t;$$

either way,

$$|f_i(x_j) - f_i(x_{j+1})| = |x_j(f_i) - x_{j+1}(f_i)| \geq 2t.$$

So for the function f_i, we have

$$\sum_{j=1}^{2^n} |f_i(x_j) - f_i(x_{j+1})| = \sum_{j=1}^{2^n} |x_j(f_i) - x_{j+1}(f_i)| \geq \sum_{j=1}^{2^n} \mathbb{I}[b_{j_i} \neq b_{j+1_i} \cdot 2t \geq \frac{2^n}{n} \cdot 2t.$$

As $\{x_j\}_{j=1}^{2^n}$ is a partition of $[0,1]$ we get

$$v \geq \sum_{j=1}^{2^n} |f_i(x_j) - f_i(x_{j+1})| \geq \frac{t2^{n+1}}{n} \geq t2^{n/2}$$

and hence

$$v/t \geq 2^{n/2}$$
$$\Rightarrow 2\log_2(v/t) \geq n.$$

■

Proof [Proof of Lemma 16] We construct a set of $n = \lfloor \log_2(v/t) \rfloor$ functions that are t-shattered by $\mathcal{F}^*$. First, we build a balanced Gray code (Flahive and Bose, 2007) with n bits, which we arrange into the rows of M. Divide the unit interval into 2^n segments and define, for each $j \in [2^n]$,

$$x_j := \frac{j}{2^n}.$$

Define the functions $f_1, \ldots, , f_{\lfloor \log_2(v/t) \rfloor}$ as follows:

$$f_i(x_j) = \begin{cases} t, & M(j,i) = 1 \\ -t, & M(j,i) = 0 \end{cases}.$$

We claim that each $f_i \in \mathcal{F}$. Since M is balanced Gray code,

$$V(M) = \frac{2^n}{n} \leq \frac{v}{t\log_2(v/t)} \leq \frac{v}{2t}.$$

Hence, for each f_i, we have

$$V(f_i) \leq 2tV(M,i) \leq 2t\frac{v}{2t} = .v$$

Next, we show that this set is shattered by $\mathcal{F}^*$. Fix the trivial offest $r_1 = ... = r_n = 0$ For every $b \in \{0,1\}^n$ there is a $j \in [2^n]$ s.t. $b = b_i$. By construction, for every $i \in [n]$, we have

$$x_j(f_i) = f_i(x_j) = \begin{cases} t \geq r_i + t, & M(j,i) = 1 \\ -t \leq r_i - t, & M(j,i) = 0 \end{cases}.$$

■

Appendix B. Sample compression for nearest-neighbor regression

Let $(\mathcal{X}, \rho)$ be a metric space and define, for $L \geq 0$, the collection $\mathcal{F}_L$ of all $f : \mathcal{X} \to [0, 1]$ satisfying

$$|f(x) - f(x')| \leq L\rho(x, x');$$

these are the L-Lipschitz functions. Gottlieb et al. (2017b) showed that

$$d_{\mathcal{F}_L}(t) = O\left(\lceil L \operatorname{diam}(X)/t \rceil^{\operatorname{ddim}(\mathcal{X})}\right),$$

where $\operatorname{diam}(\mathcal{X})$ is the diameter and ddim is the *doubling dimension*, defined therein. The proof is achieved via a packing argument, which also shows that the estimate is tight. Below we show that $d^*_{\mathcal{F}_L}(t) = \Theta(\log(M(\mathcal{X}, 2t/L)))$, where $M(\mathcal{X}, \cdot)$ is the packing number of $(\mathcal{X}, \rho)$. Applying this to the efficient nearest-neighbor regressor[3] of Gottlieb et al. (2017a), we obtain

Corollary 17 *Let $(\mathcal{X}, \rho)$ be a metric space with hypothesis class $\mathcal{F}_L$, and let $\mathcal{L}$ be a consistent, proper learner for $\mathcal{F}_L$ with target generalization error ε. Then $\mathcal{L}$ admits a compression scheme of size $O(k \log k)$, where*

$$k = O\left(D(\varepsilon) \log \frac{1}{\varepsilon} \cdot \log D(\varepsilon) \log\left(\frac{1}{\varepsilon} \log D(\varepsilon)\right)\right)$$

and

$$D(\varepsilon) = \left\lceil \frac{L \operatorname{diam}(\mathcal{X})}{\varepsilon} \right\rceil^{\operatorname{ddim}(\mathcal{X})}.$$

We now prove our estimate on the dual fat-shattering dimension of $\mathcal{F}$:

Lemma 18 *For $\mathcal{F} = \mathcal{F}_L$, $d^*_{\mathcal{F}}(t) \leq \log_2(\mathcal{M}(\mathcal{X}, 2t/L))$.*

Proof Let $\{f_1, \ldots, f_n\} \subset \mathcal{F}_L$ a set that is t-shattered by $\mathcal{F}^*_L$. For $b \neq b' \in \{0,1\}^n$, let i be the first index for which $b_i \neq b'_i$, say, $b_i = 1 \neq 0 = b'$. By shattering, there are points $x_b, x_{b'} \in \mathcal{F}^*_L$ such that $x_b(f_i) \geq r_i + t$ and $x_{b'}(f_i) \leq r_i - t$, whence

$$f_i(x_b) - f_i(x_{b'}) \geq 2t$$

and

$$L\rho(x_b, x_{b'}) \geq f_i(x_b) - f_i(x_{b'}) \geq 2t.$$

It follows that for $b \neq b' \in \{0,1\}^n$, we have $\rho(x_b, x_{b'}) \geq 2t/L$. Denoting by $M(\mathcal{X}, \varepsilon)$ the ε-packing number of $\mathcal{X}$, we get

$$2^n = |\{x_b \mid b \in \{0,1\}^n\}| \leq \mathcal{M}(\mathcal{X}, 2t/L).$$

■

3. In fact, the technical machinery in Gottlieb et al. (2017a) was aimed at achieving *approximate* Lipschitz-extension, so as to gain a considerable runtime speedup. An *exact* Lipschitz extension is much simpler to achieve. It is more computationally costly but still polynomial-time in sample size.

Lemma 19 *For $\mathcal{F} = \mathcal{F}_L$ and $t < L$, $d^*_{\mathcal{F}}(t) \geq \log_2(\mathcal{M}(\mathcal{X}, 2t/L))$.*

Proof Let $S = \{x_1, ..., x_m\} \subseteq \mathcal{X}$ be a maximal $2t/L$-packing of $\mathcal{X}$. Suppose that $c : S \to \{0,1\}^{\lfloor \log_2 m \rfloor}$ is one-to-one. Define the set of function $F = \{f_1, \ldots, f_{\lfloor \log_2(m) \rfloor}\} \subseteq \mathcal{F}_L$ by

$$f_i(x_j) = \begin{cases} t, & c(x_j)_i = 1 \\ -t, & c(x_j)_i = 0 \end{cases}.$$

For every $f \in F$ and every two points $x, x' \in S$ it holds that

$$|f(x) - f(x')| \leq 2t = L \cdot 2t/L \leq L\rho(x, x').$$

This set of functions is t-shattered by S and is of size $\lfloor \log_2 m \rfloor = \lfloor \log_2(\mathcal{M}(\mathcal{X}, 2t/L)) \rfloor$. ■